Evidence Matters

HISTORICAL METHODS, Fall 2002, Volume 35, Number 4

Missing, Biased, and Unrepresentative

The Quantitative Analysis of Multisource Biographical Data

GIDON COHEN
Politics and Contemporary History
University of Salford

Abstract. With the growth in interest in collective biography as a historical technique, many predominantly qualitative historians find themselves faced with large amounts of information. These data, collected from a variety of sources, are often highly irregular, making statistical analysis extremely problematic. Current practice is to ignore these problems and proceed with quantitative analysis suitable only for much more regular data. It is argued that a more satisfactory approach is to ascertain and directly confront the difficulties of analyzing such information. The three central problems are identified as missing data, systematic bias, and the lack of a representative sample. Using a practical example, the author explores the relationship between gender, the family, and political socialization within the Communist Party of Great Britain and shows how each of the issues can be dealt with in turn. The author first distinguishes truly missing data from "negative information," which commonly appears to be missing in historical sources. He then stratifies the data to remove systematic biases relating to the issue at hand. Finally, he divides the sample into different populations, on the basis of the sources from which individuals are known, and compares the results obtained to examine whether his conclusions appear to depend on quirks of populations contained in the sources. These ideas open a new range of sources to quantitative analysis and raise the possibility of allowing new types of evidence to count in historical inquiry.

Keywords: Britain, Communist Party, gender, missing data, prosopography

In 1998, almost 40 years after the much-heralded "quantitative revolution" in history began, the *International Review of Social History* dedicated a supplement to introducing statistical and other social scientific techniques to an audience of social historians. This in itself is indicative of the fact that, outside of economic history and historical demography, the anticipated torrent of cliometric history had "simply failed to materialise" (Griffin and van der Linden 1998, 6; cf. Reynolds 1998). Part of the reason most historians lack interest in quantitative methods lies in the distance between the methods; the subjects of interest; and, most particularly, the sources used by each group. Furthermore, because many historians are averse to the often crude application of inadequate statistical techniques to inappropriate data, they have stressed the difference between historical and scientific or social scientific inquiry. Either implicitly or explicitly they have denied the relevance of quantification in the writing of history.

Despite the apparently intractable conflict between the differing schools, historians noted for their oppositional stances found genuine agreement on many specific issues (Fogel and Elton 1983). As one conservative critic of cliometrics has pointed out, if "quantification [can be] a means rather than an end, an instrument of analysis rather than a theory about human behavior, . . . it may be accommodated in traditional history, subjected to the ordinary rules of evidence, and made part of a larger historical enterprise" (Himmelfarb 1987, 44). With this route to reconciliation in mind, the mutual suspicion between qualitative and quantitative historians has recently shown some signs of thawing with the emergence of a number of studies and projects that attempt to mix different methods and approaches (Church and Outram 1998; del Mar del Pozo Andrés and Braster 1996).

Perhaps the most notable developments have been in the field of prosopography, the study of collective biography.[1] Major prosopographical projects attempting to combine aspects of both qualitative and quantitative analysis are now under way on a wide range of historical periods from the ancient Greek world to the modern day (Kilpatrick et al. 1997; Keats-Rohan 1999; Martindale 2001; Pelteret 2001). Some of these studies are based on populations where complete or nearly complete information is available for certain categories of information (Mawdsley and White 2000). Further studies are based on a random or other probability sample of the population (Woods and Hinde 1985). The researcher can combine simple descriptive statistics, such as counts and percentages, with basic inferential tests if appropriate, to form a compelling part of an overall argument. However, only in the case of censuses on the one hand and relatively well known elite groups on the other do we have access to full historical populations or a probability sample of them.

Studies conducted on other nonelite groups, usually drawn together from a range of sources containing different types of information and levels of detail, are now increasingly common. Three particular problems are evident in the quantitative analysis of such data: missing data, systematic bias, and the dependence of the results on available sources. However, general practice has been to ignore the issues of reliability, bias, and missing data and to simply report overall counts and percentages without concern as to how they may be flawed (Clements 1997; Hillyar and McDermid 2000). Many ideas for the research that served as the basis of this article come from my collaborators Kevin Morgan and Andrew Flinn (see acknowledgments at the head of Notes). The purpose of this article was to clarify, and propose possible solutions to, the difficulties experienced when one undertakes the quantitative analysis of collective biographical data drawn from various sources and presents the general outline of solutions to the most acute problems of statistical analysis of multisource biographical data.

We hoped to present material in a manner accessible to qualitative historians, rather than to prove or even demonstrate statistically optimal techniques. We accept the widely held belief that the potency of any methodology is best displayed through the analysis of actual cases (Griffin and van der Linden 1998, 1). Thus, methodological points are introduced in an exploration of the relationship between gender, family, and political socialization within the Communist Party of Great Britain (CPGB). We estimated the percentage of communists whose parents were also members of the CPGB and examined whether female communists were more likely to have communist parents than their male counterparts. In answering these questions, we deal with each of the three difficulties in turn. First, we describe the difficulties of extracting the "negative information," such as inactivity or lack of participation, that appears in many historical sources as missing data. We propose five assumptions, of differing plausibility, to deal with the problem caused by the fact that individuals rarely explicitly declare that their parents were "not communist." Second, we address the problem of bias, noting that females had a higher propensity to have communist parents simply because on average they joined the Communist Party at a later date. We used stratification to show how one can remove this bias by comparing the males and females who joined in the same period. Combining stratification with techniques for dealing with missing data allows much more satisfactory estimates of population characteristics. Third, we address the reliability of such estimates. By comparing the results obtained from independent populations, defined by the sources used, it is possible to test whether results depend on our particular choice of sources, or alternatively whether it would be reasonable to expect to obtain similar results if completely different sources had been used. We also show that such tests can detect problems with the assumptions generated to deal with negative information.

By combining the techniques outlined, we provide estimates that do not appear to depend on our particular sources. Those estimates suggest that in some periods as many as one-third of communists came from party families. We also give evidence that female communists were more likely than their male counterparts to have had communist parents, a conclusion that again appears not restricted to particular sources. We show how working toward these conclusions can fruitfully generate ideas for further qualitative and quantitative research. The main purpose of this article was to suggest to skeptical qualitative historians that the problems of missing data, bias, and lack of a representative sample should not prevent quantitative analysis. Instead, we maintain that even with problematic data, qualitative and quantitative approaches can be usefully combined, potentially opening up a new range of sources for examination.

Gender and Communist Families

The complex relationship between gender, family, and politics has been an increasing focus of study for contemporary historians. The movement of females from the private sphere of home and family to the public sphere of politics was perhaps the central political struggle in British politics in the late nineteenth and early twentieth centuries. However, the moves toward formal voting equality did not automatically imply equality in political participation. Women remained marginalized from much political activity throughout the twentieth century. Even within socialist parties, formally committed to equality, women were usually seen as wives and daughters rather than as activists in their own right (Hunt 1996, 197–205). Thus, one central question concerns the differential impact of parental influence on the political socialization of males and females. Commentators on the British labor movement have suggested that the differences in parental influence between men and women were qualitative rather than quantitative. They have suggested that there is no significant difference in the percentages of men and women who share their parents' politics. Instead, it seems that parental influence was more important for females largely because women were more likely to recognize the formative significance of their parents' political activity (Graves 1994, 43–57). Recent studies of the family's importance as a source of recruitment within the International Communist movement have tended to follow similar lines, either neglecting the question of gender entirely or discussing it in purely qualitative terms (Cohen 1997; Kaplan and Shapiro 1998).

In this article, we examine the extent and gender balance of "party families" whose children follow their parents' politics, within the CPGB, an important component of both the International Communist and the British labor movements. We concur with the prevailing view that the party family was an important source of recruitment for both men and women and, indeed, present and check the validity of a

number of estimates of the percentages of British communists whose parents were also members of the Communist Party. However, in contrast to the assertions of other commentators, we present substantial evidence that there is a quantitative, as well as a qualitative, difference in parental influence between males and females. We argue that female communists were significantly more likely than male communists to have had communist parents. We examine the reasons for this by comparing the different patterns of recruitment in terms of gender inside and outside the family. We present evidence that, from the substantial percentages of communists joining from inside party families, men and women were recruited in roughly equal numbers. This result can be contrasted with the extreme preponderance of male recruitment outside the family. This parity of recruitment within the family, combined with the predominantly male recruitment outside the family, implies overall a greater proportion of women with communist parents.

We analyzed a database containing information on 4,248 members of the CPGB from 1920 to 1991. The database was initially conceived as a tool for storing qualitative biographical information in a structured way to facilitate data retrieval and qualitative analysis. However, with an eye to potential quantification, we designed the database to enable the flexible recall of structured information relating to many areas of an individual's life, including work, politics, family and upbringing, residence, leisure, and personal activity. It was crucial that the database record the sources used, which varied substantially, and the information they provided for each individual. One reason the world's communist parties have formed such a focus for prosopographical studies is that the bureaucratic nature of such organizations was combined with an intensive interest in the biographical details of activists. The CPGB, along with other world communist parties, required that its members complete autobiographies to attend party schools or fulfill other party functions. The result is a collection of over 3,000 autobiographies of party members now deposited in the party archives in the National Labour History Museum in Manchester. Alongside a systematic sample of these autobiographies, an oral history project of over 150 interviews was conducted. We processed further interviews and interview transcripts that did not form part of the project.[2] A range of other sources was used, including nominations to positions within the party, published and unpublished biographies and autobiographies, comments in correspondence between the CPGB and Moscow, and personal archives and correspondence.

We do not suppose that any of these sources gives access to an unvarnished truth about an individual. If autobiographies generally present the author's self to readers with a particular purpose in mind, then the Communist Party autobiography presents a particularly extreme example of this "constructed self" (Bjorklund 1998; Penettier and Pudal forthcoming). A central part of our wider aim is to study these different constructions of communist identity (Morgan, Cohen, and Flinn forthcoming). However, whereas suspicion lingers and conclusions remain necessarily somewhat tentative, information from such problematic sources can fruitfully be subjected to quantitative examination. In particular, if "selves" that are constructed for very different purposes and audiences show similar quantitative patterns, the patterns themselves imply the veracity of the self-constructions.

The reliability of our conclusions about the extent of party families within the CPGB depends on there being no systematic difference in the probability of those with communist and noncommunist parents being included on the database. One central element of the study was designed to test this premise. However, before proceeding it was necessary to remove any cases that violated this assumption. If individual communists appeared in the sources only because they were children of communists, they had to be removed from the study. In particular, certain relationships between sources and individuals applied only to those with communist parents. For example, everyone who had completed a party autobiography was a member of the CPGB, and if the autobiography provided details of children, those children inevitably had a communist parent. Similar considerations applied to a number of other sources, most notably other autobiographies and the interviews entered on the database. We identified 32 communists who were known to us only because they were referred to by their communist parents. We removed the 32 individuals from our study before the investigation began.

Missing Data and Negative Information

The vast majority of historical sources do not provide uniform and comparable pieces of information for subsequent quantitative analysis, which inevitably results in a considerable amount of missing data. There were no data on gender for a relatively small number of individuals; for others, information on background variables, such as when they joined the Communist Party, was missing. However, the greatest problem in examining the party family was caused by a lack of information about whether an individual's parents were communists. Standard statistical approaches stress the importance of determining patterns that exist in any missing data. The most widely used techniques for dealing with missing data assume that the mechanism by which the data came to be missing can safely be ignored. There are two situations in which the missing data mechanism can be ignored. First, when data are missing completely at random (MCAR)—as when each value of a variable is equally likely to be missing—if one simply deletes the missing data, no bias will be introduced. Second, the missing data mechanism can be ignored when the values for data are missing at random (MAR) *given* the observed data. Data very often contain information MAR, a classic example being a repeat survey on political attitudes

when low-income respondents are hard to trace. The bias that results from having data MAR can be addressed if one creates imputed values for each piece of missing data on the basis of patterns in the observed data. However, when certain values of a variable are more likely to be missing in a way that cannot be predicted from the observed data, the mechanism by which the data are missing cannot be ignored. Hence, this situation is described as having nonignorable (NI) missing data (King et al. 2001; Rubin 1976, 1996). The problem of nonignorable missing data is particularly common with historical data when "negative information," such as inactivity or lack of participation, is often not reported. When values such as "politically inactive" are much more likely to be missing, one cannot attempt bias removal without a specific consideration of the mechanism by which data are missing.

We attempted to distinguish parents who were members of the Communist Party from those who were not. However, only an extremely limited number of individuals explicitly stated that their parents were not in the Communist Party. The failure to declare negative information, making those with politically inactive parents less likely to report their parents' relationship to communist politics, identifies the central mechanism that causes bias in our missing data. To deal with this problem, when a direct statement of nonmembership in the Communist Party was not provided, it was necessary to infer from other information whether the parents could be assumed to be members of the Communist Party. The data were analyzed under five different assumptions that could be made about those parents who were not members of the Communist Party. The first assumption, labeled *parental political information*, included only those parents who had made an explicit declaration of political affiliation to a noncommunist party. This assumption is rather restrictive, requiring direct evidence that parents were not communists; it effectively ignores the bias that results from leaving out those with politically inactive parents. To deal with this problem, one needs some way of identifying those cases where the lack of information about parental political activity is likely to imply politically inactive parents. Given our sources, the identification can be attempted if we look at other available information about individuals and their parents.

If requiring explicit political information is too restrictive, other assumptions suffer from the opposite problem. A second assumption, labeled *database*, is far too liberal and includes all individuals in the database, assuming that if we do not have a record of parents having been party members, then they were not in fact party members. Given the complete absence of any parental information in a wide range of sources, such an assumption is clearly unwarranted. A third assumption, labeled *parental information*, is only slightly more reasonable, for it posits that where we have any parental information, all parents with no political information were politically inactive. Many sources, such as correspondence between the party in London and the Comintern in Moscow, provide limited information about the parents' names, illnesses, and deaths without any requirement that even Communist Party membership, let alone any other form of political activity, be declared.

However, in our sources, the political affiliation of parents is more frequently declared than most other facts about parents. Thus, when we have a substantial amount of information about an individual's parents, but no record of any party affiliation, it is reasonable to assume that the parents were not in the Communist Party.[3] The information about an individual's parents is considered to be substantial if there is information about both the individual's social and economic background. We labeled this fourth assumption *parental economic and social information*.[4] For illustrative purposes, we also presented results on the more relaxed assumption that parents were not in the Communist Party when there was no record of parental Communist Party membership but there was information about parental employment or union activity. We labeled this fifth assumption *parental economic information*.

Table 1 shows the percentages and gender breakdown of communists from party families under the five different assumptions. Under four of the five assumptions, there is a significant difference in the percentage of men and women with communist parents.[5] However, restricting our analysis to those whose parental political information is explicitly given, thereby effectively ignoring the negative information, we found that the difference between men and women was not statistically significant. This lack of significance appeared not only because of the reduction in overall sample size but also because we had parental political information for significantly more females than males. This finding, although requiring further research for a fully adequate explanation, is highly suggestive, especially given the overall tendency in our sources to declare political backgrounds within the labor movement more frequently than those outside the movement. In this situation, the greater percentage of men from nonpolitical backgrounds seems to indicate that women were also more likely to have had a parent active in the wider labor movement. Nevertheless, table 1 indicates that under our best assumptions there was a significant difference between the percentages of men and women with communist parents. The table also shows that, whatever the assumptions, individuals from party families made up a substantial percentage of the party's members, with the figures on the most plausible estimates being just less than 20 percent.

Stratification to Remove Bias

The use of inferential statistics in historical research has long been associated with the existence of random or other probability samples (Floud 1975, 173). However, in scientific inquiry, much work has gone into the development of

TABLE 1
Number and Percentage of Male and Female Communists with Communist Parents (CP) under Different Assumptions

Assumption	Parents: Male	Parents: Female	Total
Database			
CP			
n	70	44	114
%	2.7	4.6	3.2
Non-CP			
n	2,509	912	3,421
%	97.3	95.4	96.8
Parental information			
CP			
n	70	44	114
%	11.6	17.3	13.3
Non-CP			
n	532	211	743
%	88.4	82.7	86.7
Parental economic information			
CP			
n	70	44	114
%	13	22.9	13.9
Non-CP			
n	507	200	707
%	87.9	82	86.1
Parental economic and social information			
CP			
n	70	44	114
%	16.7	24	18.9
Non-CP			
n	349	139	488
%	83.3	76	81.1
Parental political information			
CP			
n	70	44	114
%	27.8	31	28.9
Non-CP			
n	182	98	280
%	72.2	69	71.1

Note: Database includes all on database; parental information includes all for whom we have any parental information; parental economic information includes all with economic parental information; parental economic and social information includes all with social and economic parental information; and parental political information includes all those with explicit political parental information.

statistical methods to deal with "observational" situations in which no probability sampling is possible and in which experiments cannot be done. One central need in such studies is the removal of bias that arises from background, or confounding variables. A number of different techniques have been developed to deal with this situation, on the basis of the need to ensure that only cases that are essentially alike are compared.

Perhaps the easiest to understand of the techniques to remove bias is stratification, or subclassification. One often-cited example of the potential for large data sets to suggest relationships were the observational studies on the connection between smoking and mortality. William Cochran (1968) described how large-scale observations were conducted, collecting death rates for nonsmokers, cigarette smokers, and cigar and pipe smokers. The mortality rate for cigar and pipe smokers appeared much higher than that for either cigarette smokers or nonsmokers, but there did not appear to be a significant difference in mortality between nonsmokers and cigarette smokers. However, some groups, especially pipe and cigar smokers, were notably older than other groups, such as nonsmokers and cigarette smokers. To conduct a more meaningful analysis, they had to separate the populations into strata. Then "young" nonsmokers could be compared with cigarette smokers and pipe and cigar smokers in the same age cohort, and so on. As the number of age strata increased, more bias was removed. When sufficient strata were introduced, the results of the study changed dramatically. There appeared little difference in mortality between nonsmokers and pipe and cigar smokers, but the cigarette smokers now appeared significantly more likely to die. Cochran showed that as long as there are a reasonable number of people from each group in each strata, comparisons using five or six subclasses will typically remove 90 percent or more of the bias.

Despite the raw differences in the percentages of males and females with communist parents, the presence of confounding factors means that one cannot argue that there is a direct connection between gender and the likelihood of having a communist parent. In the early years of the party, individuals were much less likely to have parents who were also members of the Communist Party. During those early years, the political socialization of parents who were supportive of left-wing politics—or even of the Communist Party itself—often took place in other socialist organizations, where they frequently remained. The early Communist Party was also a much more male-dominated organization than the post-1945 organization. The party census of 1927 declared only 17 percent females. By 1956, the percentage of females had risen to 33 percent.

To examine the relationship between gender and the propensity to have a communist parent, one needs to check how much of the larger proportion of women found with communist parents is due to the fact that, on average, women tended to join the party in a later period. The data were divided according to year of joining the Communist Party, similar to the adjustments made to compare smokers and nonsmokers with different average ages. We stratified our sample into six periods on the basis of historical judgments about the nature of the CPGB: 1920–24, the founding years of the organization; 1924–28, bolshevization; 1928–35, the (long) third period; 1935–45, the popular front and World War II; 1945–56, from the war to the Hun-

garian Revolution and Nikita Khrushchev's secret speech; and 1956–91, to the fall of the Soviet Union and the end of the Communist Party.

This process creates a series of nested contingency tables, or multiway frequency tables. Table 2 shows such a table under our most plausible assumption—that parents were not members of the CPGB when there was no record of party membership but we had parental economic and social information. The data support the idea that the differences between males and females are not due simply to the different dates at which they joined the party. In every time period, a greater percentage of females came from party families. The table also indicates considerable change in the ratios over time, indicating different patterns for males and females. In particular, the difference between males and females is least in the period from the beginning of the popular front period to the end of World War II, when the CPGB began to launch gender-specific recruitment campaigns and women were increasingly seen in the workplace. Such patterns can provide a stimulus for both qualitative analysis and further quantitative research.

More rigorous examination of such multiway frequency tables that test whether the differences in the table are likely due to chance, can be conducted using log-linear modeling. With log-linear models, one uses different amounts of information to compare observed category counts with expected values. Expected values correspond to the maximum likelihood estimates of each category count when particular pieces of information are used in the estimation. The overall difference between observed and expected values gives a measure of the goodness of fit of the model. If the goodness of fit of a model is not significantly better with a certain piece of information, it would seem that the factor is not important in explaining the distribution of the data. The significance of the improvement in goodness of fit between two models can be calculated from the change in goodness of fit and the change in the number of degrees of freedom between the models with and without pieces of information, thus giving an overall probability, corresponding approximately to the idea that the improvement in fit is due to chance alone.[6]

TABLE 2
Number and Percentage of Male and Female Communist Recruits with and without Communist Parents (CP): Parental Economic and Social Information Assumption

	Parents			
	CP		Non-CP	
Date of joining	*n*	%	*n*	%
1920–24				
Male	3	4.4	65	95.6
Female	4	19	17	81
1924–28				
Male	4	17.4	19	82.8
Female	4	30.8	9	69.2
1928–35				
Male	9	10.3	78	89.7
Female	3	16.7	15	83.3
1935–45				
Male	17	12.5	119	87.5
Female	14	16.7	70	83.3
1945–56				
Male	13	25	39	75
Female	10	43.5	13	66.5
1956–91				
Male	8	30.8	18	69.2
Female	3	50	3	50

For example, one might wonder whether the gender division of our population affects the counts we find in table 2. In other words, we need to know how good an estimate of the counts in each category is possible without knowing how many males and females there are in our population. To do this, we use log-linear models to calculate the goodness of fit of the data with and without information on gender. Essentially, the observed category counts are compared with the expected counts first with, and then without, the information as to the actual numbers of males and females. The goodness of fit of the model to the data in these two cases is then compared. The huge change in the goodness of fit shows that any model without this information is inadequate, which is, of course, implied by the fact that there are many more males than females in our population.

The primary variables in a model—in this case, gender, date of joining the party, and communist parents—are known as the main effects. Log-linear models can be used not only to investigate the impact of these main effects but also to investigate the interactions between the main effects, which is crucial. Thus, we address the question of the relationship between gender and communist parents by comparing the goodness of fit to the observed counts of expectations formed with and without the interaction between gender and parental politics.[7] If gender is unrelated to parental political affiliation, after adjustment for date of joining the CPGB, then we would expect the inclusion of the interaction between gender and parental politics to yield no significant improvement in the goodness of fit of the data.

We fitted log-linear models to our data for each of the five assumptions made about when we could assume that parents were not members of the Communist Party. As shown in table 3, under all of the four assumptions where the unstratified data had shown a relationship between gender and parental politics, the inclusion of the interaction between gender and parental politics led to a statistically significant improvement in the fit of the data.[8] The relationship between gender and the communist family was again not significant on the assumption that explicit political information is required about parents to count as not having communist parents. Although as already noted, further research is required, given the preponderance of parents in this category from labor movement backgrounds, it would seem that the influence of parental politics often took a non-

TABLE 3
Results of Log-Linear Analysis of Relationship between Gender, Communist Parents (CP), and Date of Joining the Communist Party of Great Britain

Assumption	Change in goodness of fit with removal of interaction: Logistic regression $\Delta\chi^2$	*df*	*p*
CP Parental × Gender interaction			
Database	10.675	1	.0011
Parental information	6.686	1	.0097
Parental economic information	11.788	1	.0006
Parental economic and social information	11.816	1	.0006
Parental political information	2.112	1	.1462
CP Parental × Gender × Date interaction			
Database	6.757	5	.2393
Parental information	2.687	5	.7480
Parental economic information	2.578	5	.7647
Parental economic and social information	3.213	5	.6672
Parental political information	2.974	5	.7040

Note: Database includes all on database; parental information includes all for whom we have any parental information; parental economic information includes all with economic parental information; parental economic and social information includes all with social and economic parental information; and parental political information includes all those with explicit political parental information.

institutional form. Nevertheless, overall our data pointed to a significant difference in the propensity of men and women to have communist parents, which could not be explained solely with reference to the fact that women on average joined the Communist Party later than men.

Stability of the Results

As with the vast majority of the samples available to historians, ours is a sample of convenience. Because it is not a probability sample, it is not possible to test directly the reliability of our estimates of the population as a whole using standard tools. Further biases are introduced by the problematic nature of the sources. In such a situation, one important indication of validity is the stability of our results. How different would the results have looked if we had used altogether different sources and analyzed a completely different subset of the population of interest?

One could respond to the stability question by repeating the process of collecting and analyzing data using entirely different sources and then comparing the final results with those from the existing study. The agreement of findings from separate studies would rightly be considered as powerful evidence for the veracity of the initial results. However, the results from our multisource database are, effectively, already the result of combining a large number of such processes of collecting and analyzing data about particular populations. By stratifying the data into the populations defined by sources and comparing the results, we can investigate how far the answers to our different research questions vary between different populations. If the results do not vary significantly between the different populations, it is powerful evidence in support of the results.

In general, the populations from any single source group were fairly small. Log-linear analysis rapidly loses its power when confronted with sparse cells. It is generally recommended that there should be an expectation of no more than 20 percent sparse cells, with expected counts fewer than five, and there should be no very sparse cells, with expected counts less than one. To overcome the problems of sparse data, we stratified our data into three much larger populations that describe three very different source-based subpopulations within our database. The first population is defined by a single source—all those who completed a party autobiography that came to be deposited in the Manchester archive. The second population comes from an oral history of the party and consists of all those for whom we have a recorded or transcribed interview. The final population consists of all remaining individuals. The three subpopulations are defined exclusively: when we have both a party autobiography and an interview for an individual, the assignment to the party autobiography grouping takes precedence; similarly, an assignment to an interview takes precedence over an assignment to the other sources of population.

We deal first with our conclusions about the overall percentages of individuals in our database who come from party families. To test the stability of our results, we stratified our data by source-defined population in addition to the stratification by date of joining the CPGB and whether the individual had communist parents. In effect, this created separate estimates of the percentages, with communist parents for each of our three populations. We could then examine these different estimates to see if they differed substantially between the different source groups by examining whether including the interaction between communist parents and the source population significantly improved the fit of our log-linear model.

We first discuss our "best" assumption—that parents were not in the Communist Party when parental information included both social and economic aspects along with no explicit mention of party membership. The results of the stratification by source for this assumption are shown in table 4. A summary of the results using all assumptions can be found in table 5. In general, using our best assumption, the results from different sources tend to agree with one another. When substantial differences exist between the estimated percentages, as in the 1920–24 period, the differences are based on small numbers. The results are con-

TABLE 4
Percentage of Communist Recruits with Communist Parents, by Date of Joining the Communist Party: Parental Economic and Social Information Assumption

	Sources							
	Oral history		Personnel files		Other		Total	
Date of joining	*n*	%	*n*	%	*n*	%	*n*	%
1920–24	5	40	15	0	67	7.5	89	7.9
1924–28	4	25	12	25	20	20	37	22.2
1928–35	27	14.8	30	10	49	10.2	114	11.3
1935–45	50	14	119	13.4	51	15.7	220	14.1
1945–56	21	38.1	41	29.3	14	21.4	29	30.3
1956–91	19	31.6	8	37.5	5	40	32	34.4

TABLE 5
Results of Log-Linear Analysis of Relationship between Source Population, Communist Parents (CP), and Date of Joining the Communist Party of Great Britain

	Change in goodness of fit with removal of interaction		
Assumption	Logistic regression $\Delta\chi^2$	*df*	*p*
CP Parental × Source Population interaction			
Database	74.766	2	< .0001
Parental information	8.031	2	.0180
Parental economic information	9.258	2	.0098
Parental economic and social information	1.075	2	.5841
Parental political information	.135	2	.9346
CP Parental × Source Population × Date interaction			
Database	6.594	10	.7631
Parental information	6.948	10	.7303
Parental economic information	7.683	10	.6597
Parental economic and social information	7.667	10	.6603
Parental political information	10.199	10	.4232

Note: Database includes all on database; parental information includes all for whom we have any parental information; parental economic information includes all with economic parental information; parental economic and social information includes all with social and economic parental information; and parental political information includes all those with explicit political parental information.

firmed by the log-linear analysis found in row 4 in the top panel of table 5. The interaction between sources and communist parents led to no significant improvement in the overall fit of the model. Thus, we have no reason to believe that investigations on other sources using similar methods and assumptions would yield different percentages of party members with communist parents across the time periods.

We knew that our second assumption—that all those parents for whom we had work or union information were not in the Communist Party unless we explicitly knew otherwise—was rather loose. Many examples from our sources contained fragments of such economic information, but we did not feel that the information about the parents was sufficient to allow us to assume that they were not party members. In the sense that this second assumption is sometimes violated, and those from party families are sometimes counted as having noncommunist parents, the sample proportions are not "true" measures of population percentages. Unless the probability of mistakenly assigning an individual to the non-CP group is equal in all source groups, this discrepancy will introduce systematic differences between the populations. Indeed, as shown in row 3 in the top panel of table 5, under this second assumption, the inclusion of the interaction between an individual's source population and whether they have communist parents significantly improves the log-linear model. Under this assumption, we would expect the use of different sources to yield substantially different estimates of the overall percentages with communist parents, even after accounting for the confounding effect of different dates of joining.

Table 5 shows the full results of this analysis of the relationship between source groups and parental information and date of joining the CPGB. The log-linear models tested in the top panel of the table examine whether knowledge of the sources for an individual are useful in predicting whether that individual had communist parents.[9] Overall, we found that the assumptions that we knew to be poor estimates of percentages with communist parents all produced percentages of communist parents that varied significantly from source group to source group. By contrast, when we relied on the data actually collected on parental politics or used our "best" assumption to transform missing data into negative information, the estimates of percentages with communist parents did not show dependency on the sources. These conclusions about the dependency of the data distribution on the sources are necessarily somewhat tentative. With stratification into six time periods, the data was thinly spread. However, performing the same analysis with three time periods removed the problems of sparse

cells and did not change any of our findings.[10] We thus suggest that when reasonable assumptions are used, it seems possible to obtain estimates of the percentages of communists from party families that do not depend unduly on the peculiarities of the particular sources we consulted.

Testing the consistency of our estimates of the gender breakdown was more problematic. Stratification by gender (2), time period (6), source population (3) and communist parents (2) creates 2 * 6 * 3 * 2 = 72 cells, with many expected to be sparse or very sparse. There were even many expectations of sparse cells when we used only two time periods. Therefore, although our models gave us no reason to suppose that our estimates of the percentages of men and women with communist parents depended on particular sources, we can provide no compelling evidence that studies on different sources would be likely to arrive at similar estimates for the percentages of males and females with communist parents in each time period. Nevertheless, even after stratification and with the sparse cells, the interaction between gender and communist parents remains significant. Although we cannot give reliable estimates of the extent of the difference, we retain confidence in the general conclusion that females were more likely than males to have communist parents.

The significance of these figures then lies mainly in the fact that they point to an important relationship between gender and family background in terms of recruitment to the Communist Party. Whereas the details of this relationship are beyond the scope of this article, some hints at more substantive conclusions can be indicated. Our figures demonstrate that the recruitment of the children of communists into the party was always important, in some periods accounting for as much as one-third of new members. Recruitment patterns for men and women varied, with women more likely to come from party families, which suggests that any analysis of recruitment to the CPGB needs to take account of the party family. Further, within communist families, there was no statistically significant difference between the numbers of men and women who joined the party, either with or without stratification to remove bias from date of joining. Impressive confirmation of the idea that male and female children of communists were equally likely to join the party comes from the fact that there were 22 women (48 percent) among the 46 children mentioned as party members by those who completed party autobiographies.[11]

Thus, our major substantive finding is not so much that women had a greater propensity to come from party families, for the party family appears as an important means of recruitment for men as well. Rather, we suggest that to understand the differences between male and female recruitment into the Communist Party, it is crucial to look separately at recruits made within and outside the family. Our results show that when such a distinction is made, the failures of the Communist Party to recruit females are even more apparent than the raw membership figures suggest. Whether such findings apply to other political parties is a matter for further investigation.

In addition to providing an answer to our initial research questions, the quantitative analysis raised important questions for further research, which could be pursued by either qualitative or quantitative means. Here we mention only the two most significant. First, in the process of separating negative information from missing data, we noted that there was a significant difference between men and women in their propensity to come from politically active families, quite possibly the result of the large number of children of labor movement activists who joined the Communist Party in certain periods, whereas with the party family, men and women may have been recruited in approximately equal numbers. Perhaps the informal transmission of political ideas and attitudes from parents to children within the labor movement was as important as the transmission of institutional loyalties within the true party family. Second, when we stratified the data by date of joining the CPGB, we noted that from 1935 to 1945 males and females from outside party families appeared to join the party in almost equal numbers. This development could have been the result of the gender-specific recruitment campaigns of the late 1930s or of changes in the patterns of female work during World War II. By raising such questions, our quantitative analysis shows itself to be more than just an exercise in problem solving. Gertrude Himmelfarb's (1987) challenge has been met; quantification has become a means rather than an end, a fruitful part of an ongoing process of historical investigation engaging, albeit in new and interesting ways, with the same types of historical evidence used by qualitative historians.

Conclusion: Making Evidence Count

There has been a recent increase in studies that perform statistical analysis of biographical data collected from multiple sources. Generally such work has used techniques suitable only for use on complete populations with reliable data. Statistical techniques for dealing with the particular problems of historical information from a variety of sources are, however, available, if not yet widely employed by historians. In this article, I identified three major problems faced in the analysis of multisourced prosopographical data: (1) the problem of the relationship between missing data and negative information, (2) the problems that can stem from confounding factors, and (3) the potential for unreliable results that can emerge from the use of unrepresentative populations.

To deal with each of these three problems, we first used a number of different assumptions to distinguish negative information from truly missing data. By stratifying our results, we removed the bias stemming from confounding factors. Finally, we checked the reliability of our results by examining their consistency across populations defined by different sources and successfully applied our techniques to

a study of the relationship between the family, gender, and recruitment to the Communist Party of Great Britain. There is convincing statistical evidence that the family was of central importance as a means of recruiting new communists, including estimates of the percentage of communist activists who came from party families in different time periods. It is also true that the gender composition of new recruits from within the party family was much less male dominated than the recruitment from outside the family. Thus, we demonstrated the need for any study of recruitment to the CPGB, or indeed any political organization, to study recruitment within and outside the family separately. Under our favored assumptions, none of the results appeared to depend on any of the particular sources we had used.

The primary purpose of the study was to raise awareness of the problems of missing data, bias, and unreliability, rather than to propose particular generally applicable solutions to these problems. Indeed, a number of our techniques are statistically suboptimal. Our treatment of missing data makes the strong assumption that information is missing completely at random across most of our data. The possibility that there may be other patterns in our missing data, whether identifiable from observed values of other variables or requiring explicit consideration of other missing data mechanisms, is ignored. A more extensive model for the treatment of missing data, which incorporated known biases within the remaining missing data would have been preferable (Little and Rubin 1987, 171–94; Little 1983). Furthermore, our stratification to remove bias used only one variable, whereas if we had stratified on an overall propensity score we could have controlled for any number of confounding factors (Rosenbaum and Rubin 1984).[12] Other fundamentally different approaches might have been more parsimonious. In particular, a Bayesian approach, in which initial (prior) assumptions are modified in light of the data collected, appears particularly promising.[13] Indeed, much of the discussion of missing data within the statistical literature is set in Bayesian terms (King et al. 2001; Rubin 1976; Singh 1983). Nevertheless, our treatments have the advantage of remaining comparatively simple to understand while representing a major advance in current practice.

Among the most important reasons why the quantitative revolution in historical methods floundered is the limited existence of data on complete populations, or a probability sample of such data. The development of quantitative techniques to deal with problematic data opens up a whole new range of historical sources to quantitative analysis. The qualitative historian works by collecting fragments from different sources to develop a picture of individuals, institutions, processes, and events. Although he or she may succeed in building up a number of relatively complete or at least interestingly contextualized cases, along the way a huge number of fragments of information are collected and discarded. Apparently too isolated and random to be inherently interesting, these fragments are much more common than the complete lists that provide the fodder for conventional quantitative history. Providing a mechanism for admitting these pieces of evidence, small or large, into historical debate opens a much greater prospect for reconciliation between qualitative and quantitative historians. Using quantitative techniques on qualitative data allows different types of historians to share concerns about sources and evidence to a much greater extent than they can at the present time. Techniques that deal with missing information, remove bias, and test for the impact of unrepresentative populations allow the possibility of making a new kind of evidence count in historical inquiry.

NOTES

This research was carried out as part of the CPGB Biographical Project at the University of Manchester funded by ESRC award no. R000 23 7924. This was a collaborative research project, and much of the information and ideas in this article came from Kevin Morgan and Andrew Flinn. I am grateful to Jocelyn Evans for his extensive comments on the approach to the subject and to Sarah Cohen and two anonymous referees for their comments on earlier drafts of this article.

1. L. Stone (1971) provided an accessible introduction to prosopography as the study of collective biography including an examination of the differences between qualitative and quantitative approaches. For a more technical discussion of the debates over database design for both qualitative and quantitative collective biography, see Harvey and Press (1996).

2. At the time of analysis, only 73 of the project interviews had been entered onto the database.

3. This assumption—that substantial amounts of information about a person's parents without a record of communist activity implies nonmembership in the Communist Party—relies on individuals not wishing to hide the communist affiliation of their parents. Such an assumption is not universally valid. Some individuals may wish to hide unpleasantness that stems from having communist parents. However, within communist circles, and particularly within internal party sources, a communist background was generally something to be proud of, rather than something to be hidden. Our study here is restricted to the communist children of communist parents. In general, these communists would have no reason to hide and every reason to disclose their parents' communist affiliation.

4. More precisely, a parent is assumed to be politically inactive when there is no political information but rather information on either religious upbringing—including parental religious attitudes—or formative experiences, and either parental employment or parental union.

5. One can use a chi-square test to ascertain the significance of such two-dimensional relationships. The four assumptions under discussion show a significant difference between male and females, $p < .05$.

6. Here, we restrict our attention to hierarchical log-linear models. Analysis is conducted using SPSS. For an introduction to log-linear models for historians, see Kousser, Cox, and Galenson (1982). For a more recent introduction to log-linear modeling including multiway frequency analysis using SPSS, see Tabachnick and Fidell (2001, 219–74).

7. To eliminate the bias that may result from differences in other variables, we included in both these models all the main effects and interactions between the main effects, with the exception of an interaction between gender and parental politics.

8. Under none of the assumptions did the interaction between gender, parental politics, and strata of joining date make a significant improvement to the fit of the model. Thus, our data did not point to a reliable change in the ratios of men to women with communist parents over time. Of course, in the nature of statistical tests, the possibility of a significant change in the percentages of men and women over time cannot be ruled out. A null hypothesis can be rejected or not rejected, but a failure to reject the null hypothesis does not imply that the null hypothesis should be accepted.

9. From the bottom panel of table 5, included for completeness, one can see whether knowledge of both the source group and the date of joining of

an individual is useful in predicting whether they had communist parents. The results show that such information never significantly improved the fit of the model and thus did not confound any of our conclusions.

10. The three time periods used split the database population into three equal-sized groups (1920–31, 1931–41, 1941–91). The change in goodness of fit with the removal of the interaction between source population and whether individuals have communist parents under the parental economic and social information assumption was 1.773 (2*df*), $p = .4121$, and under the parental economic information assumption was 10.274 (2*df*), $p = .0059$.

11. These children were excluded from the other results in this article except where referred to in other sources. See the previous discussion at the end of the section Gender and Communist Families (page 168).

12. Apart from date of joining the party, there are a number of confounding factors—such as differential fertility, different prominence in party activity, and geographical variations—that could bias the analysis. If these factors are of interest in themselves, they can be included in the model, although increasing the number of dimensions in a log-linear model places exponentially greater demands on the data. Thus, where the main aim is to remove the bias that results from the presence of the confounding factor, it is preferable to include it as one element in an overall propensity score.

13. I am grateful to an anonymous referee for this suggestion.

REFERENCES

Bjorklund, D. 1998. *Interpreting the self: Two hundred years of American autobiography.* Chicago: University of Chicago Press.

Church, R., and Q. Outram. 1998. *Strikes and solidarities: Coalfield conflict in Britain, 1889–1966.* Cambridge: Cambridge University Press.

Clements, B. 1997. *Bolshevik women.* Cambridge: Cambridge University Press.

Cochran, W. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24(2): 295–313.

Cohen, P. 1997. *Children of the revolution.* London: Lawrence and Wishart.

del Mar del Pozo Andrés, M., and J. Braster. 1996. Bridging the gap between quantitative and qualitative historical research: An application of multiple regression analysis and homogeneity analysis with alternating least squares. *History and Computing* 8(3): 133–45.

Floud, R. 1975. *An introduction to quantitative methods for historians.* London: Methuen.

Fogel, R. W., and G. R. Elton. 1983. *Which road to the past? Two views of history.* New Haven, Conn.: Yale University Press.

Graves, P. 1994. *Labour women: Women in British working class politics, 1918–1939.* Cambridge: Cambridge University Press.

Griffin, L., and M. van der Linden. 1998. Introduction. *International Review of Social History* 43(6): 3–8.

Harvey, C., and J. Press. 1996. *Databases in historical research: Theory, methods and applications.* Basingstoke: Macmillan.

Hillyar, A., and J. McDermid. 2000. *Revolutionary women in Russia, 1870–1917: A study in collective biography.* Manchester: Manchester University Press.

Himmelfarb, G. 1987. *The New History and the Old.* Cambridge: Belknap Press.

Hunt, K. 1996. *Equivocal feminists: The Social Democratic Federation and the woman question, 1884–1911.* Cambridge: Cambridge University Press.

Kaplan, J., and L. Shapiro, eds. 1998. *Red diapers: Growing up in the communist left.* Chicago: University of Illinois Press.

Keats-Rohan, K. S. B. 1999. Historical text archives and prosopography: The COEL database system. *History and Computing* 10: 57–72.

Kilpatrick, R., H. Short, H. Walda, and G. Waywell. 1997. The Daidalos project. *Literary and Linguistic Computing* 12(3): 177–84.

King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95(1): 49–69.

Kousser, J. M., G. W. Cox, and D. W. Galenson. 1982. Log-linear analysis of contingency tables: An introduction for historians with an application to Thernstrom on the "Floating Proletariat." *Historical Methods* 15(4): 152–69.

Little, R. 1983. The nonignorable case. In *Incomplete data in sample surveys: Vol. 2. Theory and bibliographies,* edited by W. Madow, I. Olkin, and D. Rubin, chap. 22. New York: Academic Press.

Little, R., and D. Rubin. 1987. *Statistical analysis with missing data.* New York: John Wiley.

Martindale, J. R. 2001. *Prosopography of the Byzantine Empire (641–867): The CD of the first period.* Aldershot: Ashgate.

Mawdsley, E., and S. White. 2000. *The Soviet political elite from Lenin to Gorbachev—The Central Committee and its members, 1917–1991.* Oxford: Oxford University Press.

Morgan, K., G. Cohen, and A. Flinn. Forthcoming. *People of a special mould? Communists in British society.* London: Rivers Oram.

Pelteret, D. 2001. The challenges of constructing the prosopography of Anglo-Saxon England database. *Medieval Prosopography* 22: 1–9.

Pennetier, C., and B. Pudal. Forthcoming. Communist prosopography in France: Research based on French institutional communist autobiographies. In *Agents of the revolution: Biographical and prosopographical approaches to the history of international communism,* edited by G. Cohen, K. Morgan, and A. Flinn. Oxford: Peter Lang.

Reynolds, J. F. 1998. Do historians count anymore? The status of quantitative methods in history, 1975–1995. *Historical Methods* 31(4): 141–48.

Rosenbaum, P. R., and D. B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79: 516–24.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63(3): 581–92.

———. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91:434, 473–89.

Singh, B. 1983. Bayesian approach. In *Incomplete data in sample surveys: Vol. 2. Theory and bibliographies,* edited by W. Madow, I. Olkin, and D. Rubin, chap. 22. New York: Academic Press.

Stone, L. 1971. Prosopography. *Daedalus* 100: 46–79.

Tabachnick, B. G., and L. S. Fidell. 2001. *Using multivariate statistics,* 4th ed. Boston: Allyn and Bacon.

Woods, R. I., and P. R. A. Hinde. 1985. Nuptiality and age at marriage in nineteenth-century England. *Journal of Family History* 10(2): 119–44.